The Effective Sample Size of Relative Pairs in Genetic Case-Control Association Analysis

Wentian Li[1], Yaning Yang[2], Peter K. Gregersen[1] *

1. The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, North Shore LIJ Health System, Manhasset, NY 11030, USA

2. Department of Statistics and Finance, University of Science and Technology of China Anhui 230026, Hefei, CHINA

February 9, 2008

Abstract

We address an issue in the transition from genetic linkage analysis to genetic association analysis: how to correctly account for correlations between samples obtained from a pedigree for a case-control analysis. Since correlation does not affect the mean of genotype or allele frequency estimation (it only affects the variance), we introduce the concept of "effective sample size" to account for this effect. The concept of effective sample size much simplifies the handling of complicated relationship between correlated samples. For example, for allele frequency estimation, sibpairs and parent-child pairs are equivalent to 1.5 samples, first cousins and uncle-nephew pairs are equivalent 1.6 samples, etc., without considering the affection status. For genotype frequency estimation, the effective sample size concept is perhaps less convenient because its value depends on a particular genotype and depends on the allele frequency. We present the formula for χ^2 test statistic and 95% confidence interval of odd-ratio, the two most frequently used quantities in case-control analysis, for correlated samples using the effective sample size.

*Email: wli@nslij-genetics.org,ynyang@ustc.edu.cn, peterg@nshs.edu

1 Introduction

The genetic case-control association analysis [7] is one of the most commonly used approaches in human disease gene mapping. In this analysis, researchers collect tissues or blood samples from patients as well as normal controls. The DNA molecules are extracted from the sample for genotyping, either by microsatellite markers or single-nucleotide-polymorphism (SNP) complete human genome is now technologically feasible (e.g. using chips), and have been routinely carried out. The goal of a genetic association analysis is to locate genetic markers that exhibit significantly different allele or genotype frequencies between the patient (case) and the normal (control) group.

Another competing paradigm for disease gene mapping is the genetic linkage analysis [16]. In linkage analysis, families with multiple incidents of a disease are identified, and samples are collected for genotyping. For common "complex" human diseases with unknown disease etiology and disease inheritance mode, one popular study design for linkage analysis is the non-parametric affected sibpair analysis. Many research groups, ours included[9, 10, 1], started out by collecting affected sibpairs for genetic linkage analysis, but now plan to use these patient samples in case-control association analysis also.

The transition from linkage to association analysis faces one statistical problem: for almost all statistical tests, samples are assumed to be independent. Three options are available if samples are correlated: first, the independence condition is forced by picking one sample per family; second, the independence condition is ignored while all correlated samples are included in the analysis; third, all correlated samples are included but their correlation is accounted for in the analysis. The first option loses samples thus reduces statistical power. The second option, so-called *naive* estimator, is generally unbiased, but underestimates the variance, thus reporting incorrect p-values. We focus on the third option, using an approach called "naive estimator with effective sample size", to provide a simple solution in handling correlated samples.

Although the topic of correlated samples in genetic case-control association analysis is not new (see, for example, Refs.[17, 6, 3]), our approach deviates from many other publications by not using the likelihood method. The reason for this is that we would like to provide an easy solution accessible to most practitioners of genetic analysis who may or may not have a strong mathematical background. Notice that besides the correlation between two relatives in a relative pair, most samples are still independent because they are from different pedigrees. In this context, although likelihood method can correctly account for correlation between multiple samples in a pedigree, the simple effective sample size approach is an excellent alternative that is easily understandable and can be applied using existing analysis programs.

This article is organized as follows: we reproduce the important result that correlation among samples do not bias the allele frequency estimation; then we calculate the variance of genotype

frequency when a fixed type of relative pairs is used; the concept of effective sample size is introduced; the variance calculation is extended to allele frequency estimator; the use of effective sample size in χ^2 test statistic calculation is provided; and the use of effective sample size in calculating confidence intervals of odd-ratio calculation is presented.

2 Correlation Among Samples Does Not Bias the Mean

Before examining the effect of correlation among samples, we first discuss the quantity which is *not* affected by correlation: the variable mean. Suppose all our samples consist of N_r correlated pairs: $\{x_l, y_l\}$ ($i = 1, 2, \ldots N_r$), where x_l and y_l are the variable value for individual 1 and 2 in pair l, such as the genotype indicator variable. An estimator of the mean using all samples,

$$E\left[\frac{\sum_l (x_l + y_l)}{2N_r}\right] = \sum_l \frac{E[x_l] + E[y_l]}{2N_r} = E\left[\frac{\sum_l x_l}{N_r}\right], \tag{1}$$

is identical to the estimator using one sample per pair (note that $E[x_l] = E[y_l]$). In other words, correlation between samples in a pair does not bias the mean estimator. The key of this proof is that correlation between x and y only enters as a product, whereas for the estimation of the mean, there is no cross-product term.

In a slightly more complicated situation, where the samples consist of singletons, correlated pairs, and correlated triples. We can write the singletons as $\{x_l\}$ ($l = 1, 2, \cdots N_1$), pairs as $\{x'_m, y'_m\}$ ($m = 1, 2, \cdots N_2$), and triples as $\{x''_n, y''_n, z''_n\}$ ($n = 1, 2, \cdots N_3$). The mean of the naive allele/genotype counting estimator using all N ($N = N_1 + 2N_2 + 3N_3$) samples is:

$$\begin{aligned} & E\left[\frac{\sum_l x_l + \sum_m (x'_m + y'_m) + \sum_n (x''_n + y''_n + z''_n)}{N}\right] \\ = & \frac{\sum_l E[x_l] + \sum_m (E[x'_m] + E[y'_m]) + \sum_n (E[x''_n] + E[y''_n] + E[z''_n])}{N} \\ = & E[x] = E\left[\frac{\sum_l x_l + \sum_m x'_m + \sum_n x''_n}{N_1 + N_2 + N_3}\right], \end{aligned} \tag{2}$$

again, identical to the estimator using only one uncorrelated sample per group. It can be easily seen that the same conclusion holds for other more complicated situations.

This general conclusion has directly consequence on the estimation of allele frequencies from pedigree data[4, 5]. In most pedigree analysis programs, both options are available for estimating allele frequencies: either from all individuals or from pedigree founders only.[1] It is acknowledged in Ref. [5] that "one will not go far wrong in simply using the data for all individuals, ignoring their relationship".

[1] See, e.g., the program PEDMANAGER, *http://www.broad.mit.edu/ftp/distribution/software/pedmanager/*.

If both naive estimator using correlated samples and that using independent samples lead to the same mean, why should we worry about the naive estimator? There could be these concerns: for example, if singletons and pairs are collected under different conditions, a selection bias might be introduced; and, although the means of the two estimators are the same, the variances of the two are not. For a particular dataset of finite number of samples, estimator with a larger variance may fail to reproduce the true value. We will focus on the variance calculation next.

3 Variance of Genotype Frequency Estimated from Relative Pairs

We discuss the following situation: genotype frequencies are to be estimated from a group of relative pairs of the same type, e.g., 500 sibpairs. The genetic marker is assumed to have two alleles, A and B, with allele frequency p and q, and the three possible genotypes, AA, AB, BB, with expected genotype frequencies of p^2, $2pq$ and q^2 under Hardy-Weinberg equilibrium. The genotype indicator vector[20] for the first and second relative can be written as X_i and Y_j (i, j are the genotype index). The variance of naive estimator of genotype frequency G_i (i=2,1,0 refer to AA, AB, BB genotype) is:

$$\begin{aligned} Var[\hat{G}_i] &= Var\left[\frac{\sum_l (X_{i,l} + Y_{i,l})}{2N_r}\right] = \frac{\sum_l (Var[X_{i,l}] + Var[Y_{i,l}] + 2Cov[X_{i,l}, Y_{i,l}])}{4N_r^2} \\ &= \frac{Var[X_i]}{2N_r} + \frac{Cov[X_i, Y_i]}{2N_r}. \end{aligned} \tag{3}$$

The first term in Eq.(3) is the variance of genotype frequency estimated from $2N_r$ independent samples. The second term in Eq.(3) is the extra variance due to the correlation between two relatives in a relative pair:

$$Cov[X_i, Y_i] = E(X_i, Y_i) - E(X_i)E(Y_i). \tag{4}$$

The cross-covariance matrix $Cov[X_i, Y_j]$ can be calculated by the Li-Sacks ITO matrix[12, 13, 14, 17, 8]:

$$\begin{aligned} Cov[X_i, Y_j] &= E(X_i, Y_j) - E(X_i)E(Y_j) = \sum_{k=0}^{2} P(Y_j|X_i, k)P(k)P(X_i) - P(X_i)P(Y_j) \\ &= \sum_{k=0}^{2} P(Y_j|X_i, k)\pi_k G_i - G_i G_j \end{aligned} \tag{5}$$

where k ($k = 0, 1, 2$) is the identity-by-descence (IBD) status between the two relatives, $\pi_k \equiv P(k)$ is the probability of IBD=k, G_i or G_j is the population genotype frequencies, and $P(Y_j|X_i, k)$ ($k = 0, 1, 2$) are the ITO matrices[12, 13, 14, 8]:

IBD=2 IBD=1 IBD=0

	Pr(IBD=2)=π_2	Pr(IBD=1)=π_1	Pr(IBD=0)=π_0	kinship-coefficient ϕ
parent-child	0	1	0	1/4
sibs	1/4	1/2	1/4	1/4
half-sibs	0	1/2	1/2	1/8
uncle/aunt-nephew/niece	0	1/2	1/2	1/8
first cousins	0	1/4	3/4	1/16
second cousins	0	1/16	15/16	1/64

Table 1: IBD probability for several types of relative pair.

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad T = \begin{pmatrix} p & q & 0 \\ p/2 & 1/2 & q/2 \\ 0 & p & q \end{pmatrix}, \quad O = \begin{pmatrix} p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \end{pmatrix}. \tag{6}$$

It can be shown that the variance of the naive genotype frequency estimator is:

$$\begin{aligned} Var[\hat{G}_2] &= \frac{p^2(1-p^2)}{2N_r} + \frac{p^2(\pi_2 + p\pi_1 + p^2\pi_0)}{2N_r} - \frac{p^4}{2N_r} \\ Var[\hat{G}_1] &= \frac{2pq(1-2pq)}{2N_r} + \frac{2pq(\pi_2 + 0.5\pi_1 + 2pq\pi_0)}{2N_r} - \frac{(2pq)^2}{2N_r} \\ Var[\hat{G}_0] &= \frac{q^2(1-q^2)}{2N_r} + \frac{q^2(\pi_2 + q\pi_1 + q^2\pi_0)}{2N_r} - \frac{q^4}{2N_r} \end{aligned} \tag{7}$$

For two unrelated individuals, $\pi_2 = \pi_1 = 0$, $\pi_0 = 1$, the second and the third terms in Eq.(7) cancel, and only the first term remains.

The expected IBD probability between two relatives of different types is easily available [11] and some of them are included in Table 1. Take sibpairs for example, π_2=1/4, π_1 =1/2, π_0 =1/4, and the variances of genotype frequencies by Eq.(7) are:

$$\begin{aligned} Var[\hat{G}_2]_{\text{sib}} &= \frac{p^2(1-p^2)}{2N_r} + \frac{p^2(1+3p)(1-p)}{8N_r} \\ Var[\hat{G}_1]_{\text{sib}} &= \frac{2pq(1-2pq)}{2N_r} + \frac{pq(1-3pq)}{2N_r} \\ Var[\hat{G}_0]_{\text{sib}} &= \frac{q^2(1-q^2)}{2N_r} + \frac{q^2(1+3q)(1-q)}{8N_r} \end{aligned} \tag{8}$$

Variance of genotype frequency in other relative pairs can be derively similarly by inserting the IBD probability frmo Table 1 to Eq.(7). For parent-child pairs and uncle-nephew pairs, the results are:

$$Var[\hat{G}_2]_{\text{par-child}} = \frac{p^2(1-p^2)}{2N_r} + \frac{p^3q}{2N_r}$$

$$
\begin{aligned}
Var[\hat{G}_1]_{\mathrm{par-child}} &= \frac{2pq(1-2pq)}{2N_r} + \frac{pq(1-4pq)}{2N_r} \\
Var[\hat{G}_0]_{\mathrm{par-child}} &= \frac{q^2(1-q^2)}{2N_r} + \frac{q^3 p}{2N_r}
\end{aligned} \tag{9}
$$

$$
\begin{aligned}
Var[\hat{G}_2]_{\mathrm{uncle-ne}} &= \frac{p^2(1-p^2)}{2N_r} + \frac{p^3 q/2}{2N_r} \\
Var[\hat{G}_1]_{\mathrm{uncle-ne}} &= \frac{2pq(1-2pq)}{2N_r} + \frac{pq(1-4pq)/2}{2N_r} \\
Var[\hat{G}_0]_{\mathrm{uncle-ne}} &= \frac{q^2(1-q^2)}{2N_r} + \frac{q^3 p/2}{2N_r}
\end{aligned} \tag{10}
$$

The second terms in Eq.(8, 9, 10) are always non-negative (note: $1-3pq = (p-q)^2+pq \geq 0$, and $1-4pq = (p-q)^2 \geq 0$), indicating that correlation among these relatives always increases the variance. We can also show that the second terms in Eq.(8) are always smaller than the first term, indicating that using two relatives always reduce the variance when genotype frequencies are estimated from one relative randomly selected from a relative pair.

4 The Effective Sample Size to Account for the Increase of Variance of Genotype Frequencies

In Eqs.(7,8), we have seen variances of genotype frequencies are higher than those of uncorrelated samples (with the same number of individuals). For binomial distribution, variance is inversely proportional to sample size. Consequently, an increase of the variance due to correlation can be accounted for by a *decrease* of the "effective" sample size. We define effective sample size N_e as the number of independent samples that lead to the same level of variance as calculated from the correlated samples.

By this definition, and by examining the analytic expression of variances in Eqs.(7,8), it is clear that for genotype frequencies, the effective sample size depends on the genotype, as well as on allele frequency p. Fig.1 shows variances of three genotype frequencies in the sibpairs, parent-child-pairs, and uncle-nephew-pairs (upper plots, solid lines), and the corresponding variances for independent samples ($2N_r$ individuals) (in dashed lines). The ratio of the two variances is shown in the lower part of Fig.1.

To get a sense of a typical sample size reduction, we carry out two averaging processes. The first is to average the three variance ratios with weight of $p^2, 2pq, q^2$ for the three genotypes, at a fixed value of p. This average is represented by a black line in Fig.1 (lower part). The second average is to average over p (x-axis in Fig.1), which leads to $\approx 1.414, 1.328, 1.164$ for the sibpairs, parent-child-pairs, and uncle-nephew-pairs. Since by definition these variance ratios

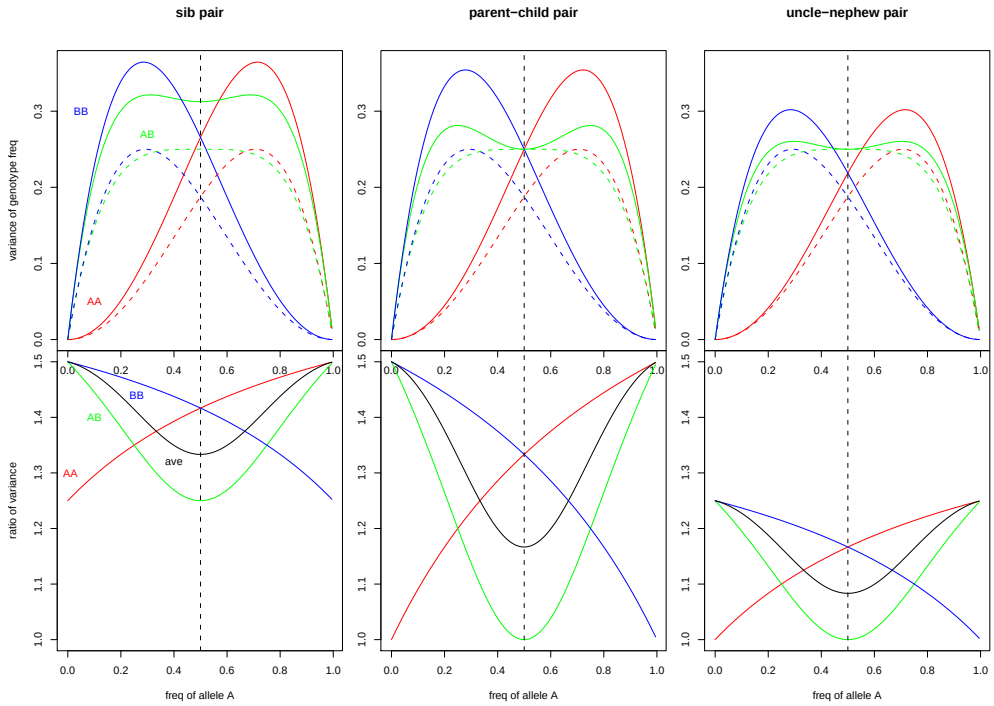

Figure 1: (Upper) Variances of genotype frequencies for the sibpairs, parent-child-pairs, and uncle-nephew-pairs (solid lines). Three colors indicate different genotypes (red: AA, green:AB, blue:BB). Dashes lines are the corresponding variances calculated from uncorrelated samples. (Lower) Variance ratios (solid lines over dashed lines) for three different genotypes. The black line is the weighted average of the variance ratio of three different genotypes.

are equal to $2N_r/(2N_e)$, we conclude that sibpairs, parent-child-pairs, and uncle-nephew-pairs are equivalent to $\approx 1.4, 1.5, 1.7$ effective samples, in order to account for the increase of variance in genotype frequencies.

5 Variance of Allele Frequency Estimated from Relative Pairs

Denote by C_l and D_l the number of allele A in the first and the second relative in relative pair l, it is clear that $C = iX_i$ and $D = jY_j$ ($i, j = 2, 1, 0$ are AA, AB, and BB genotypes). The variance of allele frequency estimated from all relatives in relative pairs is

$$\begin{aligned} Var[\hat{p}_A] &= Var\left[\frac{\sum_l (C_l + D_l)}{2 \cdot 2N_r}\right] = \frac{N_r Var[C] + N_r Var[D]}{16N_r^2} + \frac{N_r 2Cov[C,D]}{16N_r^2} \\ &= \frac{Var[C]}{8N_r} + \frac{Cov[C,D]}{8N_r} = \frac{2pq}{8N_r} + \frac{Cov[C,D]}{8N_r} \end{aligned} \tag{11}$$

Using the ITO matrix and the relationship $C = iX_i$ and $D = jY_j$, it can be shown that[2]:

$$\begin{aligned} Var[\hat{p}_A] &= \frac{pq}{4N_r} + \frac{(4p^2+2pq)\pi_2 + (4p^2+pq)\pi_1 + 4p^2\pi_0 - 4p^2}{8N_r} \\ &= \frac{pq}{4N_r} + \frac{pq(\pi_1+2\pi_2)}{8N_r} = \frac{pq}{4N_r} + \frac{4pq\phi}{8N_r} = \frac{pq(1+2\phi)}{4N_r} \end{aligned} \quad (12)$$

where ϕ is the kinship coefficient, defined as the probability that an allele selected randomly from one person is IBD with a similarly randomly selected allele from another person [15, 11]. In absence of inbreeding, the kinship coefficient of IBD=2 relative pairs (e.g. identical twins) is 1/2, and that of IBD=1 relative pairs (e.g., parent-child pairs) is 1/4. Combining the two, we have $\phi = \frac{1}{2}\pi_2 + \frac{1}{4}\pi_1$.

The kinship coefficients of the common relative pairs are listed in Table 1, and the increase of variance of allele frequency due to related individuals can be obtained immediately: for sib and parent-child pairs, the variance is increased from 1 to 1+2/4=1.5 (in the unit of $pq/(4N_r)$), and for uncle/aunt-nephew/niece and half-sib pairs, the variance is increased from 1 to 1+2/8=1.25. By definition, these variance ratio is inversely proportional to the effective sample size over the actual sample size, so the effective sample size for a sibpair (or parent-child-pair) is 4/3 ≈ 1.33, and that for a uncle-nephew-pair (or half-sib-pair) is 8/5 $= 1.6$. Interestingly, these numbers are similar but not identical to the averaged effective sample size based on variances of genotype frequencies.

6 Using Effective Sample Size in Chi-square Tests

The effective sample size method can be applied to modify Pearson's χ^2 test originally designed for independent samples. Denote the allele counts in a 2-by-2 table as $N_{\text{A,case}}, N_{\text{B,case}}, N_{\text{A,con}}, N_{\text{B,con}}$, the Pearson's chi-square test statistic is:

$$X^2 = \frac{(N_{\text{A,case}}N_{\text{B,con}} - N_{\text{B,case}}N_{\text{A,con}})^2(N_{\text{A,case}} + N_{\text{B,case}} + N_{\text{A,con}} + N_{\text{B,con}})}{(N_{\text{A,case}} + N_{\text{B,case}})(N_{\text{A,con}} + N_{\text{B,con}})(N_{\text{A,case}} + N_{\text{A,con}})(N_{\text{B,case}} + N_{\text{B,con}})}. \quad (13)$$

If control samples are independent, whereas case samples are from N_r relative pairs of the same type (e.g. all sibpairs) (so $N_{\text{A,case}} + N_{\text{B,case}} = 2 \cdot 2N_r$), we can reduce the apparent allele counts $N_{\text{A,case}}, N_{\text{B,case}}$ by a factor of $\alpha = 1/(1+2\phi)$ (e.g., for sibpairs, 2/3) when the affection status is ignored. The corrected X^2 test statistic is:

$$X_e^2 = \frac{\alpha(N_{\text{A,case}}N_{\text{B,con}} - N_{\text{B,case}}N_{\text{A,con}})^2(\alpha N_{\text{A,case}} + \alpha N_{\text{B,case}} + N_{\text{A,con}} + N_{\text{B,con}})}{(N_{\text{A,case}} + N_{\text{B,case}})(N_{\text{A,con}} + N_{\text{B,con}})(\alpha N_{\text{A,case}} + N_{\text{A,con}})(\alpha N_{\text{B,case}} + N_{\text{B,con}})} \quad (14)$$

It can be shown that $X_e^2/X^2 < 1$ (but $> \alpha$), and the reduced X^2 test statistic value leads to a larger (less significant) p-value.

[2] The same formula was derived in Ref. [5] by exhaustively counting the mating types.

	TT	TC	CC	$N_{\text{singleton}}$	N_{sibpair}	T	C	N	N_e
case	21	241	578	86	377	283	1397	1680	1177.33
control	9	143	774	926	0	161	1691	1852	1852

Table 2: Genotype counts of a SNP in gene PTPN22

In a more general situation, the patient group may consist of singletons, sib-pairs, uncle-nephew-pairs, etc. The overall effective sample size can be obtained by applying the reduction to specific relative pairs. For example, 60 singletons, 10 sibpairs and 10 uncle-nephew pairs lead to effective sample size of (ignoring the affection status) $60 + 20 \times 2/3 + 20 \times 4/5 = 89.33$, or a reduction of $\alpha = 0.8933$.

7 Using Effective Sample Size in Calculation of Confidence Interval of Odd Ratio

Another common task in case-control analysis is to estimate the 95% confidence interval (CI) of odd-ratio (OR). The estimation of OR is straightforward: $\hat{\theta} = N_{\text{A,case}} N_{\text{B,con}} / (N_{\text{A,con}} N_{\text{B,case}})$. As shown by Woolf[21], the 95% CI of OR can be approximated as:

$$[l, u] = [e^{\log\hat{\theta} - 1.96\hat{\sigma}(\log\hat{\theta})}, e^{\log\hat{\theta} + 1.96\hat{\sigma}(\log\hat{\theta})}]. \tag{15}$$

where the standard error of the logarithm of $\hat{\theta}$ can be written in four allele counts:

$$\hat{\sigma}(\log\hat{\theta}) = \left(\frac{1}{N_{\text{A,case}}} + \frac{1}{N_{\text{B,case}}} + \frac{1}{N_{\text{A,con}}} + \frac{1}{N_{\text{B,con}}} \right)^{1/2}. \tag{16}$$

Similar to our discussion in the last section, when relative pairs are involved in the patient samples, the apparent sample size is reduced to the effective sample size by a factor of α, and the above formula is modified to

$$\hat{\sigma}_e(\log\hat{\theta}) = \left(\frac{1}{\alpha N_{\text{A,case}}} + \frac{1}{\alpha N_{\text{B,case}}} + \frac{1}{N_{\text{A,con}}} + \frac{1}{N_{\text{B,con}}} \right)^{1/2}. \tag{17}$$

Since $\alpha < 1$, the standard error $\hat{\theta}$ is increased, the consequently, the 95% CI of OR is expanded.

8 Illustration by a Real Dataset

The data we use here is taken from Ref.[2] (replication study, all sibs option, in table 1 of Ref.[2]) collected for the NARAC project[9, 10, 1]. The control samples are independent, whereas case samples consist of both singletons and sibpairs. Discount the 377 sibpairs (377 $\times$ 4 alleles) by a

factor of 2/3, the effective sample size for the case group is reduced from 1680 to 1177.33, with a reduction rate of $\alpha = 0.7008$.

Using Eq.(13), the original X^2 is 53.26, corresponding to p-value of 2.9×10^{-13}. With the effective sample size, the corrected X^2 is 45.75, corresponding to p-value of 1.3×10^{-11}. As for the 95% CI of odds-ratio, the original interval is [1.73, 2.61]. After reduce the allele counts in case group by a factor of α, the 95% CI of odds-tatio becomes [1.70, 2.66]. In general, for a very significant test result, the correction to p-value due to correlation in relative pairs does not render the result insignificant, but the exact number for p-value is slightly changed. Same is true for 95% CI of odds-ratios. Only for border-line significant result, introducing correlation among samples may change the result to be insignificant[17].

9 Discussion

The effective sample size framework may not be applied to situations where case and control samples are taken from the same pedigree[19]. The problem is that case and control groups are still distinct entities in Pearson's χ^2 test. Fortunately, using both affected and unaffected individuals from the same pedigree in a case-control analysis, though remains a theoretical possibility, is not commonly practiced.

Sibship of more than two sibs, or a set of multiple relatives from the same family, can be treated in a similar way as sibpairs or relative pairs. Take three siblings for example, the variance of genotype/allele frequency will contain three covariance terms, one for each sibpair. The effective sample size and sample size reduction can be determined accordingly.

Adding affection status has a different effect from correlation among samples. For a group of affected samples, the genotype frequencies of the linked disease gene are altered to $G_{\mathrm{i,aff}} = (f_i/K)G_i$ ($i = 2, 1, 0$), where f_2, f_1, f_0 are the penetrance of the three genotypes (AA, AB, BB); and the allele frequency for the mutant allele is changed to $p_{\mathrm{A,aff}} = (p/K)(f_2p + f_1q)$. As can be seen from Fig.1, a change of allele frequency may indeed change the variance ratio as well as the corresponding effective sample size. But this effect is small due to the limited range of ratio ratio in Fig.1. We plan to carry out future analytic and simulation analysis to confirm this conjecture.

An unexplored idea is to use identical-by-state (IBS) status to calculate the joint probability: $P(X_i, Y_j) = \sum_k P(Y_j|X_i, k)Pr(IBS = k)P(X_i)$. If this indeed can be done, then IBD status is not essential to the discussion in this way: it is simply a tool in calculating the covariance.

In conclusion, the effective sample size provides a complete solution to account for the correlation between relatives if samples consist of relative pairs in a case-control study. We believe this approach is simpler and more intuitive than other mathematically sophisticated methods.

Acknowledgments

This work was supported by the NIH grant N01-AR12256. We would like to thank Prof. Chris Amos and Prof. Yongchao Ge for discussions.

References

[1] C.I. Amos, et al., High-density SNP analysis of 642 caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33. *Gene and Immunity*, 7:277-286, 2006.

[2] A.B. Begovich, et al., A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Human Genet.*, 75:330-337, 2004.

[3] S. Biedermann, E. Nagel, A. Munk, H. Holzmann, A. Steland. Tests in a case-control design including relatives. *Scandinavian J. Stat.*, to appear, 2006.

[4] M. Boehnke. Allele frequency estimation from data on relatives. *Am. J. Hum. Genet.*, 48:22–25, 1991.

[5] K.W. Broman. Estimation of allele frequencies with data on sibships. *Genet. Epid.*, 20:307–315, 2001. erratum, 23:465-466, 2002.

[6] S.R. Browning, et al., Case-control single-marker and haplotypic association analysis of pedigree data. *Genet. Epidemiology*, 28:110-122, 2005.

[7] H.J. Cordell, D.G. Clayton. Genetic association studies. *Lancet*, 366:1121-1131, 2005.

[8] F. Dai and D.E. Weeks. Ordered genotypes: an extended ITO method and a general formula for genetic covariance *Am. J. Hum. Genet.*, 78:1035-1045, 2006.

[9] D. Jawaheer, et al., A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. *Am. J. Human Genet.*, 68:927-936, 2001.

[10] D. Jawaheer, et al., Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families. *Arthritis & Rheumatism*, 48:906-916, 2003.

[11] K. Lange. *Mathematical and Statistical Methods for Genetic Analysis.* (Springer, New York, 1997)

[12] C.C. Li and L. Sacks. The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics*, 10:347-360, 1954.

[13] W. Li. A revised Li-Sacks formula for calculating the probability of identity-by-descent proportion. *Am. J. Hum. Genet.*, s63:A297, 1998.

[14] W. Li and J. Reich. A complete enumeration and classification of two-locus disease models. Hum. Hered. , 50:334-349, 2000.

[15] G. Malécot. *Les Mathématiques de l'Hérédité* (Masson et Cie, Paris, 1948).

[16] J. Ott. *Analysis of Human Genetic Linkage*, third edition, (The Johns Hopkins University Press, Baltimore, 1999).

[17] S.L. Slager, D.J. Schaid. Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects. *Am. J. Hum. Genet.*,68:1457-1462, 2001.

[18] B.K. Surez. The affected sib pair IBD distribution for HLA-linked disease susceptibility genes. *Tissue Antigens*, 12:87-93, 1978.

[19] J. Teng, N. Risch. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. individual genotyping. *Genome Res.*, 9:234-241, 1999.

[20] B.S. Weir. *Genetic Data Analysis II.* (Sinauer, Sunderland, MA, 1996).

[21] B. Woolf. On estimating the relationship between blood group and disease *Ann. Human Genet.*, 19:251-253, 1955.